Digital Trust & Safety Partnership

DTSP Submission to the United Nations on the Global Digital Compact

March 8, 2024

The Digital Trust & Safety Partnership (DTSP) welcomes the opportunity to provide feedback on the proposed structural elements for a Global Digital Compact.

Our partnership brings together providers of diverse digital products and services around shared commitments to trust and safety, and a framework of best practices and assessments grounded in the experience of practitioners. Current DTSP partners are listed on our website.[1]

DTSP welcomes the inclusion within Commitment 2 the area of action to "Advance digital trust and safety, including specific measures to protect women, children, youth and persons in vulnerable situations against harms." However, such a focus should build upon existing efforts within industry, in the trust and safety field, as well as complementary efforts by other stakeholders. Moreover, such efforts should be anchored in the protection and promotion of human rights as the introduction to Section 3 states.

We are pleased to share more information about our approach below.

1) A framework of industry best practices enables providers of diverse products and services to align around shared commitments to trust and safety online

DTSP launched in February 2021 to mature and professionalize trust and safety, the field of industry professionals dedicated to a safer and more trustworthy internet. The partnership is committed to developing, using, and promoting industry best practices, reviewed through internal and independent third-party assessments, to ensure consumer trust and safety when using digital services.

Each organization in the DTSP is guided by its own values, product aims, and experiences with user behavior. Each brings digital tools, and blended machine and human processes to make decisions about how to enable a broad range of human expression and activity, while working to mitigate as much risk as possible by identifying and preventing harmful content or conduct. Despite the individual approaches, DTSP members agree on the need for a shared framework of best practices to help raise the bar on trust and safety operations across industry and create meaningful and robust standards for assessment.

[1] https://dtspartnership.org/

All participating DTSP partners agree on five fundamental commitments that a digital service should make to promote a safer and more trustworthy internet.[2] Also known as the Digital Trust & Safety Partnership Best Practices Framework, we regard the following to be industry best practices:

- **Product Development:** Identify, evaluate, and adjust for content- and conduct-related risks in product development.
- **Product Governance:** Adopt explainable processes for product governance, including which team is responsible for creating rules, and how rules are evolved.
- **Product Enforcement:** Conduct enforcement operations to implement product governance.
- **Product Improvement:** Assess and improve processes associated with content- and conduct-related risks.
- **Product Transparency:** Ensure that relevant trust and safety policies are published to the public, and report periodically to the public and other stakeholders regarding actions taken.

DTSP regards the five overarching Commitments as representing the necessary steps taken by partner companies to identify and address harmful content and conduct while preserving free expression and other rights, including internationally recognized human rights standards, as well as the social and economic value of digital services.

These five commitments are underpinned by 35 specific (but non-comprehensive) best practices, which provide concrete though non-exclusive examples of the variety of activities and processes that organizations may have in place to mitigate risks from harmful content and conduct, depending on their particular product and risk model.

All DTSP partners embrace the Commitments, but each company is responsible for implementing a combination of the best practices that mitigate content- and conduct-related risks and ensure adherence to these commitments.

2) Rigorous evaluation using a thorough but flexible framework allows us to understand where practices are more mature and where they are in need of improvement

2a) The Safe Framework provides a methodology for internal and external assessment

Since we launched, DTSP has also developed and published our methodology for assessing the practices of partner companies. The Safe Framework establishes rigor and consistency for the assessment process, while also providing flexibility across the diverse products and services our members provide.[3] The Safe Framework examines the people, processes, and technology that contribute to managing content- and conduct-related risks for participating companies.

[2] https://dtspartnership.org/best-practices/
[3] https://dtsp.wpengine.com/wp-content/uploads/2021/12/DTSP_Safe_Framework.pdf

Using a risk-based approach, the depth of assessment is determined by evaluating the size and scale of the organization, as well as the potential impact of its product or service. Impact is based on user volume and the presence of product features or complexity that introduce potential risks.

The assessment is designed to help organizations understand how DTSP practices will help them manage content- and conduct-related risks. The outcome of the assessment will help organizations better understand the current state of their capabilities and their dependencies with respect to people, processes, and technologies. The resulting understanding can inform internal investment decisions and external engagements with policymakers, users, and civil society.

2b) Inaugural assessments identified successes and areas of improvement across our partnership

In 2022, founding partner companies conducted initial internal Safe Framework assessments, culminating in the inaugural Safe Assessments report.[4] The report revealed key findings on trust and safety practices that are well-developed across the companies and those areas that are still evolving. Many companies reported a mature state of development for core content moderation practices, such as building teams responsible for establishing, updating and enforcing the rules of their services. Several of the practices rated as least mature, on the other hand, related to engaging the perspectives of users and external organizations in trust and safety including human rights groups and academic researchers' concern engaging the perspectives of users and external organizations. The area where most improvement is expected in the future is integrating trust and safety into product development from the beginning.

3) Guiding principles and best practices for age assurance provide an example of how our framework can be applied to identify examples of practices to protect young people online

DTSP underscores that a holistic commitment to trust and safety, including robust implementation of the DTSP Best Practices Framework, is an important means of protecting young people online. As part of this overarching framework, individually and collectively, partner companies are also taking other steps that are specifically focused on designing safe age-appropriate experiences. Our recently released report, Age Assurance: Guiding Principles and Best Practices, provides examples of the specific types of practices that companies are using for this purpose.[5]

A variety of age assurance approaches exist, including age verification based on review of identity documents or parental consent; age estimation based on inferences made from user data, physical characteristics, or other measures; and self-declaration by the user.

Several challenges emerge in developing age assurance approaches. Key characteristics that digital service providers look to incorporate in developing these approaches include:

[4] https://dtsp.wpengine.com/wp-content/uploads/2022/07/DTSP_Report_Safe_Assessments.pdf
[5] https://dtspartnership.org/wp-content/uploads/2023/09/DTSP_Age-Assurance-Best-Practices.pdf

- **Effective:** Having confidence that a user is a given age allows a digital service provider to provide them with an age-appropriate experience, in a way that is accurate and hard to circumvent.
- **Accessible, inclusive and equitable:** Age assurance should not result in inequitable outcomes for a given user, and the complexity of processes should not overly burden users in ways that discourage appropriate use of a service.
- **Privacy-preserving and data-protecting:** Protecting a user's privacy, especially a young user's privacy, demands adherence to key privacy principles including data minimization, as well as implementation of security measures to protect data.
- **Affordable:** Implementation costs must be reasonable and proportionate.
- **Risk-appropriate:** The approaches deployed should be proportionate to the risks associated with underage access to a given service, as well as the risks of inaccurate determinations regarding age.

Incorporating each characteristic comes with trade-offs, and there is no one-size-fits-all solution. Highly accurate age assurance methods may depend on collection of new personal data such as facial imagery or government-issued ID. Some methods that may be economical may have the consequence of creating inequities among the user base. And each service and even feature may present a different risk profile for younger users; for example, features that are designed to facilitate users meeting in real life pose a very different set of risks than services that provide access to different types of content.

We identify five guiding principles and then note how companies have used these principles to develop example best practices for age assurance. Of course, the specific practices that services use may vary by digital product or feature and evolve with both the challenges faced and advances made in age assurance technologies.

The five guiding principles are:

1. Identify, evaluate and adjust for risks to youth to inform proportionate age assurance methods, as part of implementing safety-by-design.
2. Account for risks to user privacy and data protection as part of development, implementation, and ongoing assessment of age assurance approaches.
3. Ensure assurance approaches are broadly inclusive and accessible to all users, regardless of age, socioeconomic status, race, or other characteristics.
4. Conduct layered enforcement operations to implement age assurance approaches.
5. Ensure that relevant age assurance policies and practices are transparent to the public, and report periodically to the public and other stakeholders regarding actions taken.

4) Agreed baseline definitions of key terminology will help facilitate informed dialogue between industry, policymakers, regulators, and the wider public

Finally, our partnership emphasizes the importance of common understanding of key terminology across organizations responsible for trust and safety online, including industry actors as well as partners in government and civil society. To this end, DTSP released an inaugural edition of its Trust & Safety Glossary

of Terms.[6] This is the first industry effort by technology companies, representing various products, sizes, and business models, to develop a common trust and safety lexicon.

The Trust & Safety Glossary of Terms consists of more than 100 terms across four categories:

- content concepts and policies;
- common types of abuse;
- enforcement practices; and
- trust and safety technology.

The glossary has been updated to incorporate valuable input received from academic organizations, industry partners, regulators, and other global stakeholders during the public consultation held earlier this year. In September 2023, DTSP staff and advisors David Sullivan, Farzaneh Badiei and Alex Feerst authored a commentary for the Journal of Online Trust and Safety that provides more detail on why and how this glossary was developed.[7]

We appreciate the opportunity to provide comments based on the progress our partnership has made thus far. We look forward to the opportunity to share our experience with UN member states and other stakeholders, and to learn from the experience of others as we collectively work to make the internet safer and more trustworthy.

[6] https://dtspartnership.org/wp-content/uploads/2023/07/DTSP_Trust-Safety-Glossary_July-2023.pdf
[7] https://www.tsjournal.org/index.php/jots/article/view/147